PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-268675
(43)Date of publication of application : 20.09.2002

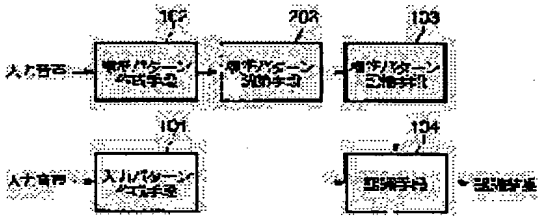
(51)Int.Cl. G10L 15/14
G10L 15/06

(21)Application number : 2001-070108
(22)Date of filing : 13.03.2001

(71)Applicant : NEC CORP
(72)Inventor : SHINODA KOICHI

(54) VOICE RECOGNITION DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a voice recognition device which effectively adjust an element distribution number at a high speed as to a probability model using a mixed distribution.
SOLUTION: This voice recognition device uses the probability model using the mixed distribution and composed of a standard pattern storage means 103 which holds a standard pattern, a recognition means 104 which inputs a voice and outputs the recognition result by using the standard pattern, a standard pattern generating means 102 which inputs a voice for learning and generates the standard pattern, and a standard pattern adjusting means 203 which adjusts the element distribution number of the mixed distribution of the standard pattern. Consequently, tree structures of element distributions are generated by states in voice recognition using a hidden Markov model having the mixed Gaussian distribution as an output probability distribution and the element distribution number of the respective states are adjusted by using an information amount reference.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's

decision of rejection]
[Date of extinction of right]

CLAIMS

[Claim(s)]

[Claim 1] A voice recognition unit characterized by to provide a standard-pattern storage means are a voice recognition unit using a probability model using mixed distribution, and hold a standard pattern, a recognition means consider voice as an input and output a recognition result using a standard pattern, a standard-pattern creation means consider voice for study as an input and create a standard pattern, and a standard-pattern accommodation means adjust the number of element distribution of mixed distribution of a standard pattern.

[Claim 2] A voice recognition unit characterized by to provide a standard-pattern storage means are a voice recognition unit using a probability model using mixed distribution, and hold a standard pattern, a recognition means consider voice as an input and output a recognition result using a standard pattern, a standard-pattern correction means consider voice for adaptation-izing as an input, and correct a standard pattern, and a standard-pattern accommodation means adjust the number of element distribution of mixed distribution of a standard pattern.

[Claim 3] A voice recognition unit according to claim 1 or 2 characterized by providing a standard-pattern accommodation means which consists of a tree structure creation means to create the tree structure of element distribution, and an element distribution selection means to choose distribution by considering study data as an input.

[Claim 4] A voice recognition unit according to claim 1 or 2 with which said standard-pattern accommodation means is characterized by providing a minimax distribution selection means to use a minimax method for selection of element distribution.

[Claim 5] A voice recognition unit according to claim 3 with which said element distribution selection means is characterized by using the study amount of data corresponding to each element distribution as a selection criterion in selection of element distribution.

[Claim 6] A voice recognition unit according to claim 3 with which said element distribution selection means is characterized by using the description length minimum criteria as a selection criterion in selection of element distribution.

[Claim 7] A voice recognition unit according to claim 3 with which said element distribution selection means is characterized by using the Akaike information criterion as a selection criterion in selection of element distribution.

[Claim 8] A voice recognition unit according to claim 3 with which said tree structure creation means is

characterized by using divergence as a distance between distribution in selection of element distribution.
[Claim 9] A voice recognition unit according to claim 3 with which said tree structure creation means is characterized by using likelihood to study data as a distance between distribution.
[Claim 10] A voice recognition unit given in either of claim 1 to claims 9 characterized by using a hidden Markov model as a probability model using mixed distribution.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[The technical field to which invention belongs] This invention relates to the voice recognition unit using the hidden Markov model using especially mixed Gaussian distribution (or gauss mixed distribution) as output probability distribution about the standard-pattern creation method in the pattern recognition which used mixed distribution.

[0002]

[Description of the Prior Art] In recent years, research on recognition by the machine of a voice pattern is done, Many methods are proposed. As technique typical in this There is a method using a hidden Markov model (HMM). And it is the purpose about the ability to recognize also in whose voice as a voice recognition system using HMM. The recognition system of the unspecified speaker carried out is studied and developed briskly.

[0003] Hereafter, HMM is taken for an example and a voice recognition system is explained based on drawing 2 . Utterance of a speaker inputted into the voice recognition unit, Every unit called the frame which is inputted into the input configuration creation means 101, and has a certain time amount length through processes, such as an AD translation and a voice analysis It is changed into the time series of a feature vector. This feature vector Time series is called an input configuration here. Moreover, the length of a frame is usual. Ten to 100ms It is a degree. And a feature vector is what extracted the characteristic quantity of the voice spectrum in the time of day, and is usually 100 dimensions from ten dimensions.

[0004] HMM is memorized by the standard-pattern storage means 103. HMM is one of the models of the audio information source, and can learn the parameter using a speaker's voice. Explanation of a recognition means describes HMM in detail. Here, HMM is usually prepared for every recognition unit. Moreover, a phoneme is taken for an example as a recognition unit here. For example, in an unspecified speaker recognition system, the unspecified speaker HMM learned beforehand, using utterance of many speakers as HMM of a standard-pattern storage means is used.

[0005] And the word "HMM" is used with the recognition means 104. An input configuration is recognized. Here, HMM is the model of the audio information source, and it serves as statistics and a probable model in order to cope with various fluctuation of a voice pattern. Moreover, detailed explanation of HMM is RABINA, JUANGU work, and the Furui translation. 102-187 pages of "the base (below) of speech recognition" and the NTT advance technology (1995) (following, reference 1) can be started.

[0006] HMM of each phoneme usually consists of state transitions of one to ten conditions, and meantime,

respectively. Usually, ****** and a final state are defined, for every unit time amount, a symbol is outputted from each condition and a state transition is performed. The voice of each phoneme is expressed as time series of the symbol outputted from HMM between the state transitions from ****** to a final state.

[0007] The appearance probability of a symbol is defined as each condition, and transition probability is defined as each transition between conditions. A transition probability parameter is a parameter for expressing the time fluctuation of a voice pattern. A output probability parameter expresses fluctuation of the voice impersonation of a voice pattern. Utterance is from the model by setting the probability of ****** to a certain value, and imposing appearance probability and transition probability for every state transition. The probability to generate can be searched for.

[0008] On the contrary, when utterance is observed and it assumes that it generated from a certain HMM, the probability of occurrence can be calculated. HMM which will search for the probability of occurrence and will serve as max in each HMM if HMM is prepared to each recognition candidate and utterance is inputted by the speech recognition by HMM by this It is decided that it will be a generation source and it considers as a recognition result with the recognition candidate corresponding to the HMM.

[0009] Although a output probability parameter has a discrete probability-distribution expression and a continuous-probability-distribution expression, a continuation probability expression is taken for an example here. In a continuous-probability-distribution expression, mixed Gaussian distribution, i.e., the distribution which added two or more Gaussian distribution with weight, is often used. In the following examples, a output probability is taken as mixed gauss continuous probability distribution. And a output probability parameter, a transition probability parameter, and two or more Gaussian distribution Parameters, such as a weighting factor, give the study voice corresponding to a model, and are beforehand learned by the algorithm called a BAUMU Welch algorithm.

[0010] For example, the case where 1000 words are now made applicable to recognition is assumed. That is, the case where the recognition candidate of 1000 words is asked for the correct answer of one word is assumed. First, when recognizing a word, it is HMM of each phoneme. It connects and HMM of a recognition candidate word is created. In 1000 word recognition, the word for 1000 words "HMM" is created. The following (1) type shows input configuration O expressed as time series of a feature vector.

[Equation 1]

$$O=o_1,o_2,o_3, \ldots ,o_t, \ldots ,o_T \qquad \cdots (1)$$

Here, T is the total frame number of an input configuration.

[0011] Moreover, recognition candidate word It is referred to as W1, W2, ··WN. N here shows the number of recognition candidate words. And each recognition candidate word "Wn" Matching between the word "HMM", and input configuration O is performed as follows. In future explanation, as long as there is no necessity, Subscript n is omitted. First, in the word "HMM", from Condition j, the mean vector of cim and each element Gaussian distribution is set to muim, and a covariance matrix is set [the transition probability to Condition i] to sigmaim for the mixed weight of aji and output probability distribution. Here, input time of day, and i and j express the condition of HMM, and, as for m, t expresses a mixed element number. The next recurrence formula count about forward probability at (i) is performed.

[0012] This forward probability at (i) is a partial observation sequence. It is the probability which outputs

o1, o2, ··, ot, and exists in Condition i in time of day t.

[Equation 2]

$$\alpha_t(i)=\pi_i \quad (i=1,2 \ldots ,I) \qquad \cdots (2)$$

[Equation 3]

$$\alpha_{t+1}(i)=\sum_i \alpha_t(j)\ a_{ij}b_i(o_t) \quad (i=1, \ldots ,I;t=1, \ldots T) \qquad \cdots (3)$$

Here, pii is the probability for an initial state to be i.

[0013] Moreover, bi (ot) in (3) types is defined by (4) and (5) types which are shown below.

[Equation 4]

$$b_i(o_t)=\sum_m \lambda_{im} N(o_t;\mu_{im},\Sigma_{im}) \qquad \cdots (4)$$

[Equation 5]

$$N(o_t;\mu_{im},\Sigma_{im})=(2\pi)^{-K/2}|\Sigma_{im}|^{-1/2}\exp(-(\mu_{im}-o_t)\Sigma^{-1}(\mu_{im}-o_t)/2) \qquad \cdots (5)$$

In this (5) type, K is the number of dimension of an input frame and the mean vector.

[0014] Moreover, the likelihood to the input pattern to the word "Wn" is called for by (6) types shown below.

[Equation 6]

$$P^n(X)=\alpha_T(I) \qquad \cdots (6)$$

In this (6) type, I is a final state.

[0015] This processing is performed about each word model, and the recognition result word to input configuration X "Wn (here, the upper part of n has **; hat by following the (7) formula)" is called for by (7) types shown below.

[Equation 7]

$$\hat{n} = \mathrm{argmax}_n P^n(X) \qquad \cdots (7)$$

And this recognition result word "Wn" is sent to the recognition result output section. The recognition result output section processes sending the control instruction corresponding to an output or a recognition result for a recognition result on a screen at another equipment etc.

[0016] Next, the standard-pattern creation means 102 is explained. In unspecified speaker recognition,

the standard-pattern creation means 102 accumulates utterance of many prior speakers, and presumes a parameter using the utterance. First, backward probability is introduced by the following (8) and (9) types.

[Equation 8]

$$\beta_T(i)=1 \quad (i=1, \ldots, N) \qquad \cdots (8)$$

[Equation 9]

$$\beta_t(i)=\sum_{j=1}^{N} a_{ij} b_j(o_{t+1})\beta(j), \qquad \cdots (9)$$

$$(t=T-1, T-2, \ldots, 1 \quad i=1, \ldots, N)$$

(9) When, as for betat (i) in a formula, time of day t and Condition i are given It is the probability of the partial observation sequence from time of day t+1 to termination.,

[0017] And when the observation sequence O is given using forward probability and backward probability, the probability which exists in Condition i at time of day t is given by (10) types shown below.

[Equation 10]

$$\gamma(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{I} \alpha_t(i)\beta_t(i)} \qquad \cdots (10)$$

Moreover, it exists in Condition i at time of day t, and exists in Condition j at time of day t+1. Probability is given by the following (11) types.

[Equation 11]

$$\zeta_{t(i,j)}=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{I}\sum_{j=1}^{I} \alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \qquad \cdots (11)$$

Moreover, in the case of mixed Gaussian distribution, it is time of day t. The k-th of a state number i The probability (occupancy frequency) which exists in a mixed element is given by the following (12) types.

[Equation 12]

$$\gamma'(i,k)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{I} \alpha_t(i)\beta_t(i)} \cdot \frac{c_{ik}N(o_t,\mu_{ik},\Sigma_{ik})}{\sum_{m=1}^{M} c_{im}N(o_t,\mu_{im},\Sigma_{im})} \qquad \cdots (12)$$

[0018] It is based on the above calculated value, The estimate of pi, a, mu, sigma, and c is given by the following (13) - (17) types.

[Equation 13]

$$\overline{\pi_i}=\gamma_1(i) \qquad \cdots (13)$$

[Equation 14]

$$\overline{a_{ij}} = \frac{\Sigma_{t=1}^{T-1} \zeta_t^{i,j}}{\Sigma_{t=1}^{T-1} \gamma_t^i} \quad \cdots (14)$$

[Equation 15]

$$\overline{c_{jk}} = \frac{\Sigma_{t=1}^{T} \gamma_t'(j,k)}{\Sigma_{t=1}^{T} \gamma_t(j)} \quad \cdots (15)$$

[Equation 16]

$$\overline{\mu_{jk}} = \frac{\Sigma_{t=1}^{T} \gamma_t'(j,k) o_t}{\Sigma_{t=1}^{T} \gamma_t'(j,k)} \quad \cdots (16)$$

[Equation 17]

$$\overline{\Sigma_{jk}} = \frac{\Sigma_{t=1}^{T} \gamma_t'(j,k)(o_t - \mu_{jk})(o_t - \mu_{jk})^t}{\Sigma_{t=1}^{T} \gamma_t'(j,k)} \quad \cdots (17)$$

[0019] With a BAUMU-Welch algorithm, a parameter is updated based on such estimate and the repeat of newly presuming estimate is further performed using the updated parameter. And it is proved that the probability to recognize an observation sequence becomes large for every repeat. In the above, the case where HMM was used was taken for the example and the conventional voice recognition unit was explained.

[0020]

[Problem(s) to be Solved by the Invention] Now, as mentioned above, discrete distribution and continuous distribution are shown in an output probability-distribution expression. And in discrete distribution and continuous distribution, mixed Gaussian distribution is well used also continuous distribution, especially in it. The reason this mixed Gaussian distribution is used is because the engine performance of an output probability-distribution expression is excellent.

[0021] Here, when using mixed Gaussian distribution (it considers as mixed distribution hereafter), there is no clear indicator which magnitude the number of element distribution should be made. Usually, the number of element distribution for every condition presupposes that it is fixed covering all conditions, some numbers of element distribution are tried on HMM of mixed distribution, and procedure of choosing the number of element distribution with the highest engine performance in it is performed.

[0022] However, it is expected that the required number of element distribution changes with conditions. For example, when it has many unnecessary element distribution, increase of the computational complexity for calculating the probability of element distribution will be caused. Moreover, in a condition with few counts of an appearance, the engine performance to strange data to which fault study will not be carried out in process of parameter estimation may deteriorate. Therefore, as for the number of element distribution in each condition of mixed distribution HMM, being optimized for every condition is desirable.

[0023] And the simplest method of optimizing the number of element distribution for every condition, It is the method of choosing the number of element distribution to which the number of element distribution is changed for every condition, a recognition experiment is conducted, and the recognition engine performance becomes high for every condition. However, it is almost unable for the number of conditions of HMM to increase very much with 1000 to 10000 on the whole, and to usually optimize the number of

element distribution for every condition in respect of computational complexity.

[0024] It is in this invention having been made under such a background and offering a high speed and the voice recognition unit performed effectively for accommodation of the number of element distribution in the probability model using mixed distribution.

[0025]

[Means for Solving the Problem] A voice recognition unit of this invention is a voice recognition unit using a probability model which used mixed distribution, and is characterized by to provide a standard-pattern storage means hold a standard pattern, a recognition means consider voice as an input and output a recognition result using a standard pattern, a standard-pattern creation means consider voice for study as an input and create a standard pattern, and a standard-pattern accommodation means adjust the number of element distribution of mixed distribution of a standard pattern.

[0026] The voice recognition unit of this invention is a voice recognition unit using a probability model which used mixed distribution, and is characterized by to provide a standard-pattern storage means hold a standard pattern, a recognition means considers voice as an input and output a recognition result using a standard pattern, a standard-pattern correction means consider voice for adaptation-izing as an input, and correct a standard pattern, and a standard-pattern accommodation means adjust the number of element distribution of mixed distribution of a standard pattern.

[0027] A voice recognition unit of this invention is characterized by providing a standard-pattern accommodation means which consists of a tree structure creation means to create the tree structure of element distribution, and an element distribution selection means to choose distribution by considering study data as an input. It is characterized by a voice recognition unit of this invention possessing a minimax distribution selection means by which said standard-pattern accommodation means uses a minimax method for selection of element distribution.

[0028] It is characterized by using the study amount of data corresponding to each element distribution for a voice recognition unit of this invention as a selection criterion in selection of element distribution of said element distribution selection means. It is characterized by using the description length minimum criteria for a voice recognition unit of this invention as a selection criterion in selection of element distribution of said element distribution selection means. It is characterized by using the Akaike information criterion for a voice recognition unit of this invention as a selection criterion in selection of element distribution of said element distribution selection means.

[0029] It is characterized by using divergence for a voice recognition unit of this invention as a distance between distribution in selection of element distribution of said tree structure creation means. It is characterized by using likelihood [as opposed to study data in said tree structure creation means] for a voice recognition unit of this invention as a distance between distribution. A voice recognition unit of this invention is characterized by using a hidden Markov model as a probability model which used mixed distribution.

[0030]

[Embodiment of the Invention] Hereafter, the operation gestalt of this invention is explained with reference to a drawing. Drawing 1 is the block diagram showing a basing-on 1 operation gestalt of this invention configuration. A different point from the conventional example of drawing 2 is inserting the standard-pattern creation means 203 between the standard-pattern creation means 102 and the standard-pattern storage means 103. In the block of the voice recognition unit of drawing 1 , to the same

configuration (the input configuration creation means 102, the standard-pattern creation means 101, the standard-pattern storage means 103, recognition means 104) as the block of the voice recognition unit of drawing 2 , the same sign is attached and detailed explanation is omitted.

[0031] In this drawing, the input configuration creation means 102 creates an input configuration from the input voice (sound signal which the speaker generated) inputted. Moreover, the standard-pattern creation means 102 creates a standard pattern, as explanation of the conventional example described. The standard-pattern accommodation means 203 is the created standard pattern. The number of element distribution is changed. The standard-pattern storage means 103 memorizes the created standard pattern, and the recognition means 205 recognizes the inputted voice using a standard pattern, and it outputs a recognition result.

[0032] Actuation of the standard-pattern accommodation means 203 added to 1 operation gestalt in this invention below is explained to details. The problem of optimization of the number of element distribution in the condition of a hidden Markov model (HMM) can be regarded as the problem which chooses the optimal probability model to the given data. selection of this probability model -- setting -- the past -- various information criteria have been proposed.

[0033] With 1 operation gestalt, how to optimize the number of distribution using MDL (description length min) which is one of them is considered. First, the criteria of Above MDL are explained here. It is proved that the description length minimum (Minimum Description Length; MDL) criteria are effective in the problem which chooses the optimal probability model from research of the latest information theory and a computation theory-learning theory to data.

[0034] The description length minimum criteria are for example, South Korean ****** and "the mathematical principle of the Iwanami lecture applied mathematics 11, information, and agreement-izing", It is explained to 249 pages - 275 pages of Iwanami Shoten (1994) (following, reference 2). It is about data easy [if possible] and moreover given like AIC (Akaike Information Criterion; Akaike information criterion) etc. It is one of the criteria which embodied the idea that the model which can be expressed was a good model.

[0035] MDL criteria are data s=s1, --, a model that gives the smallest description length to sN in a probability model i= 1, --, I. They are the criteria used as the optimal model. Here, the description length IMDL to probability-model i (i) is given by the following (18) formulas.

[Equation 18]

$$I_{MDL(i)} = -\log P_{\hat{\theta}(i)}(x^N) + \frac{\alpha_i}{2}\log N + \log I \qquad \text{... (18)}$$

here -- alphai -- number of dimension (number of a free parameter) of Model i ** theta (i) was presumed using Data XN. It is the maximum likelihood estimator of free parameter theta(i) = (theta 1 (i), --, thetaalphai (i)) of Model i.

[0036] a logarithm [on the above-mentioned (18) formula and as opposed to data in the 1st term] -- it is the amount which attached minus sign to likelihood (it is hereafter described as likelihood), the 2nd term is an amount showing the complexity of a model, and the 3rd term is description length which requires in order to choose Model i. Thus, the likelihood to data becomes large and it follows, so that a model is more complicated. The value of the 1st term decreases. On the other hand, if a model becomes complicated, since a free number of parameters will increase, the value of the 2nd term increases. Thus, the relation of

a trade-off between the 1st term and the 2nd term is, and it is expected that the description length IMDL (i) will take the minimum value with the model which has suitable complexity.

[0037] And the number optimization algorithm of element distribution for every condition using these MDL criteria is as follows. First, mixed Gaussian distribution HMM using study data is learned in the usual procedure. Under the present circumstances, the number of element distribution presupposes that it is fixed covering all conditions, and learns the increase of the number of element distribution, and HMM carried out to the number considered to be a maximum. Moreover, occupancy frequency gamma't (i, k) for every element distribution is saved in process of study. It is the subscript of element distribution [in / i and / in k / a condition] here. [the subscript of a condition]

[0038] Next, the standard-pattern adjustment means 203 optimizes the number of element distribution in each condition. In addition, this point is made to explain only one condition i, and omits the subscript i of a condition. The standard-pattern adjustment means 203 performs the same processing also to other conditions. First, the standard-pattern adjustment means 203 creates the tree structure of element distribution in every condition with an internal tree structure creation means. Here, the root is one distribution and a leaf is each element distribution.

[0039] Although various methods for creating the tree structure of element distribution at this time can be considered, a binary tree is created here using a k-means algorithm. Moreover, cull back divergence is used as a distance during each element distribution (distance between distribution). This cull back divergence is easily calculable from the value of the average and covariance of Gaussian distribution. The tree structure creation method of this element distribution is indicated by patent No. 002531073 and the above-mentioned reference 2 at details.

[0040] Next, the standard-pattern adjustment means 203 asks for distribution of distribution (node distribution) of each node of the above-mentioned tree structure. Here, distribution of each node distribution is called for from the occupancy frequency and Gaussian distribution parameter of element distribution of all leaves to govern. Now, the set of the node distribution which divides this tree structure up and down is called "a cut." Although a large number [the number of these cuts], each cut becomes one probability model in that condition. Here, it considers asking for the optimal cut using MDL criteria.

[0041] For example, the description length to a certain cut U is calculated as follows. Node distribution which constitutes Cut U here It is referred to as S1 and --SM. Here, M is the number of the node distribution in Cut U. Thereby, likelihood L (Sm) to the distribution Sm of data can be approximated like (19) and (20) types which are shown below.

[Equation 19]

$$L(s_m)=\sum_{t=1}^{T}\sum_{s\in S_m}\log(N(o_t,\mu_{Sm},\Sigma_{Sm}))\gamma_{t(s)}$$

$$=-\frac{1}{2}(\log((2\pi)^K|\Sigma|+K)\Gamma_{(Sm)} \qquad \cdots (19)$$

[0042] It sets at an above-mentioned (19) ceremony, and is [Equation 20].

$$\Gamma_{(Sm)}=\sum_{t=1}^{T}\sum_{s\in S_m}\gamma_{t(s)} \qquad \cdots (20)$$

It comes out, and it is, s is all leaf distribution under Distribution Sm, and K is the mean vector used as a share standard pattern, and the number of dimension of distribution. Moreover, in (19) types, muSm and

sigmaSm are the mean vector and distributions in Distribution Sm, respectively.

[0043] By using the result mentioned above can describe description length I (U) to Cut U like the following (21) types.

[Equation 21]

$$I(U)=\sum_{m=1}^{M} L(Sm)+\frac{1}{2}\cdot 2KM\log\sum_{m=1}^{M}\Gamma(Sm)$$

$$=\frac{1}{2}\sum_{m=1}^{M}\Gamma(Sm)\log(|\Sigma Sm|)+KM\log V+\frac{K}{2}(1+\log(2\pi))\cdot V \quad \cdots (21)$$

It is here and is [Equation 22].

$$V=\sum_{m=1}^{M}\Gamma(Sm) \quad \cdots (22)$$

It is, and it comes out and it is [it is an amount equivalent to the frame number of all the data corresponding to U, and / this V is not based on the method of division, but] constant value.

[0044] And the standard-pattern adjustment means 203 is related with all possible cuts, and is description length. l (U) is calculated and the cut U with smallest I (U) is chosen. At this time, the class of possible division, i.e., the number of Cuts U, usually increases very much. Then, the computational complexity at the time of selection of Cut U is saved by using the following algorithms. Hereafter, the number optimization of element distribution of a certain condition p is described.

[0045] First, the node (joint) to Condition p is created. Here, this node is called a root node. The distributed parameter of a root node is presumed from all the data samples corresponding to all element distribution corresponding to this condition. for example, the tree structure -- a binary tree -- it is -- distribution of a root node -- S0, the -- two -- a ** -- a child node -- distribution -- S -- one -- S -- two -- ** -- having carried out -- the time -- a parent node -- from -- a child node -- having developed -- the time -- description -- merit -- change -- a part -- the following -- (-- 23 --) -- a formula -- describing -- having .

[Equation 23]

$$\Delta=I(S_1,S_2)-I(S_0)$$

$$=\frac{1}{2}(\Gamma(S_1)\log|\Sigma_{S_1}|+\Gamma(S_2)\log|\Sigma_{S_2}|-\Gamma(S_0)\log|\Sigma_{S_0}|+K\log V \quad \cdots (23)$$

[0046] For example, the standard-pattern adjustment means 203 develops a parent node, when it is delta< 0, and on the other hand, when it is delta> 0, it does not develop a parent node. moreover -- the time of developing -- further -- child nodes S1 and S2 -- processing in which it judges whether change of the description length when developing to the child node as well as the processing mentioned above is calculated, and it develops about each is repeated. And when expansion of all nodes finishes, it means that the set of the node of the end of the expansion is cut, and the node distribution was chosen as element distribution. And mixed Gaussian distribution HMM which has only the distribution chosen anew as element distribution is created, and procedure which learns the element distribution with the data in study anew is performed.

[0047] The above is explanation of the voice recognition unit of 1 operation gestalt shown in drawing 1 . Here, although the hidden Markov model (HMM) was made into the example and explained, also when a

model is mixed Gaussian distribution, it can apply easily. This supports invention of claim 10. Moreover, although explanation of 1 operation gestalt mentioned above explained sound model study, also in case speaker adaptation which corrects a standard pattern using little utterance of a user is performed, it is possible to use the data for speaker adaptation and to adjust the number of element distribution. In this case, as a configuration of the voice recognition unit of invention, instead of a standard-pattern creation means, a standard-pattern correction means is used and the voice of the same speaker as the speaker who uses for the input configuration creation means for recognition is used for the input voice to this standard-pattern correction means.

[0048] Moreover, in the voice recognition unit of 1 operation gestalt mentioned above, although the accommodation means of the number of element distribution by the tree structure was explained, accommodation by the minimax distribution selection means using a minimax method can also be performed as follows. Hereafter, one condition is explained. First, more than the count (X time) in study data, the set of distribution which appeared is set to A and distribution which is not so is set to B. All of distribution belonging to A, distribution belonging to B, and the distance of ** are calculated, and the distance from distribution of nearest A removes the largest distribution among distribution of B.

[0049] Next, distribution with the largest distance from distribution of nearest A among distribution of B other than the distribution is removed. This procedure is repeated until the number of distribution turns into the number of the minimum distribution defined beforehand. And when not becoming smaller than the number of the minimum distribution (that is, the number of distribution of B is small), above-mentioned processing is suspended at the time. The above corresponds to invention of claim 4.

[0050] Moreover, in 1 operation gestalt, although MDL criteria were used for selection of a node, it is also possible to use an amount-of-data threshold. That is, the set of distribution nearest to a leaf is considered as a cut among a certain distribution beyond a threshold with the amount of data. The above corresponds to invention of claim 5.

[0051] Furthermore, in 1 operation gestalt, although only the case where MDL criteria were used as an information criterion was explained, when the Akaike information criterion (AIC) is used, or when other similar information criteria are used, it can apply easily. The above corresponds to invention of claim 7.

[0052] In addition, in 1 operation gestalt, although divergence was used as a distance during distribution, the increment of the likelihood when sharing distribution can also be used as a distance value. The above corresponds to invention of claim 9.

[0053] As mentioned above, although 1 operation gestalt of this invention has been explained in full detail with reference to a drawing, a concrete configuration is not restricted to this operation gestalt, and even if the design change of the range which does not deviate from the summary of this invention etc. occurs, it is included in this invention.

[0054]

[Effect of the Invention] In the pattern recognition using mixed Gaussian distribution using the parameter accommodation means which was newly added according to the voice recognition unit of this invention By adjusting to the number of element distribution to which the recognition engine performance becomes about the number of element distribution of an audio standard pattern, and becomes high for every condition of optimization, i.e., HMM, about the number of element distribution for every condition of HMM Unnecessary element distribution can be excluded, the deterioration to the strange voice data based on fault study will be prevented, and it becomes possible to perform highly

efficient speech recognition.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1] It is the block diagram showing the configuration of the voice recognition unit by 1 operation gestalt of this invention.

[Drawing 2] It is the block diagram showing the configuration of the voice recognition unit by the conventional example.

[Description of Notations]

101 Input Configuration Creation Means

102 Standard-Pattern Creation Means

103 Standard-Pattern Storage Means

104 Recognition Means

203 Standard-Pattern Accommodation Means

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2002－268675
（P2002－268675A）
(43)公開日　平成14年９月20日(2002.9.20)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　15/14		Ｇ１０Ｌ　3/00	５３５Ｃ	５Ｄ０１５
15/06			５２１Ｃ	
			５２１Ｄ	
			５２１Ｇ	

審査請求　未請求　請求項の数10　ＯＬ　（全　８　頁）

(21)出願番号　特願2001－70108(P2001－70108)

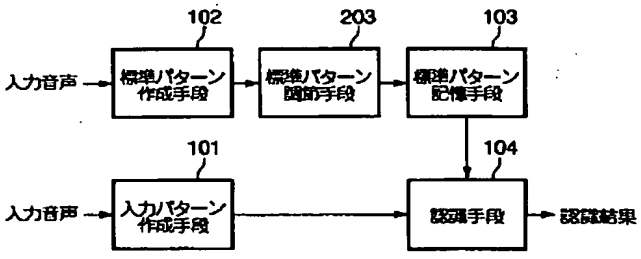
(22)出願日　平成13年３月13日(2001.3.13)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　篠田　浩一
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100108578
弁理士　高橋　詔男　（外３名）

Ｆターム(参考)　5D015 GG01 HH00

(54)【発明の名称】　音声認識装置

(57)【要約】
【課題】　混合分布を用いた確率モデルにおいて、要素分布数の調節を高速、かつ効果的に行う音声認識装置を提供する。
【解決手段】　本発明の音声認識装置は、混合分布を用いた確率モデルを用いる音声認識装置であって、標準パターンを保持する標準パターン記憶手段１０３と、音声を入力とし標準パターンを用いて認識結果を出力する認識手段１０４と、学習用音声を入力とし標準パターンを作成する標準パターン作成手段１０２と、標準パターンの混合分布の要素分布数を調節する標準パターン調節手段２０３とで構成されている。これにより、混合ガウス分布を出力確率分布としてもつ 隠れマルコフモデルを用いた音声認識において、状態ごとに要素分布の木構造を作成し、各状態の要素分布数を情報量基準を用いて調節する。



1

【特許請求の範囲】

【請求項１】　混合分布を用いた確率モデルを用いる音声認識装置であって、
標準パターンを保持する標準パターン記憶手段と、
音声を入力とし標準パターンを用いて認識結果を出力する認識手段と、
学習用音声を入力とし標準パターンを作成する標準パターン作成手段と、
標準パターンの混合分布の要素分布数を調節する標準パターン調節手段とを具備することを特徴とする音声認識装置。

【請求項２】　混合分布を用いた確率モデルを用いる音声認識装置であって、
標準パターンを保持する標準パターン記憶手段と、
音声を入力とし標準パターンを用いて認識結果を出力する認識手段と、
適応化用音声を入力とし標準パターンを修正する標準パターン修正手段と、
標準パターンの混合分布の要素分布数を調節する標準パターン調節手段とを具備することを特徴とする音声認識装置。

【請求項３】　要素分布の木構造を作成する木構造作成手段と、学習データを入力として分布を選択する要素分布選択手段とから構成される標準パターン調節手段を具備することを特徴とする請求項１または請求項２に記載の音声認識装置。

【請求項４】　前記標準パターン調節手段が、要素分布の選択にミニマックス法を用いるミニマックス分布選択手段を具備することを特徴とする請求項１または請求項２に記載の音声認識装置。

【請求項５】　前記要素分布選択手段が、要素分布の選択において各要素分布に対応する学習データ量を選択基準として用いることを特徴とする請求項３に記載の音声認識装置。

【請求項６】　前記要素分布選択手段が、要素分布の選択において、記述長最小基準を選択基準として用いることを特徴とする請求項３に記載の音声認識装置。

【請求項７】　前記要素分布選択手段が、要素分布の選択において、赤池情報量基準を選択基準として用いることを特徴とする請求項３に記載の音声認識装置。

【請求項８】　前記木構造作成手段が、要素分布の選択において、ダイバージェンスを分布間距離として用いることを特徴とする請求項３に記載の音声認識装置。

【請求項９】　前記木構造作成手段が、学習データに対する尤度を分布間距離として用いることを特徴とする請求項３に記載の音声認識装置。

【請求項１０】　混合分布を用いた確率モデルとして、隠れマルコフモデルを用いることを特徴とする請求項１から請求項９のいずれかに記載の音声認識装置。

【発明の詳細な説明】

2

【０００１】

【発明の属する技術分野】本発明は、混合分布を用いたパターン認識における標準パターン作成方法に関し、特に混合ガウス分布（またはガウス混合分布）を出力確率分布として用いた隠れマルコフモデルを用いた音声認識装置に関する。

【０００２】

【従来の技術】近年、音声パターンの機械による認識に関する研究が行われ、 数々の方法が提案されている。この中で代表的な手法としては 隠れマルコフモデル(HMM)を用いた方法がある。 そして、HMMを用いた音声認識システムとして、誰の声でも認識できることを目的 とした不特定話者の認識システムが盛んに研究・開発されている。

【０００３】以下、HMMを例にとり、音声認識システムについて図２に基づき説明する。音声認識装置に入力された話者の発声は、 入力パターン作成手段101に入力され、AD変換、音声分析などの過程を経て、ある時間長をもつフレームと呼ばれる単位ごとの 特徴ベクトルの時系列に変換される。この特徴ベクトルの 時系列を、ここでは入力パターンと呼ぶ。また、フレームの長さは通常 10msから100ms 程度である。そして、特徴ベクトルは、その時刻における音声スペクトルの特徴量を抽出したもので、通常10次元から100次元である。

【０００４】標準パターン記憶手段103にはHMMが記憶されている。 HMMは音声の情報源のモデルの1つであり、話者の音声を用いてそのパラメータを学習することができる。　HMMについては認識手段の説明で詳しく述べる。ここで、HMMは通常各認識単位ごとに用意される。また、ここでは、認識単位として音素を例にとる。例えば、不特定話者認識システムでは、標準パターン記憶手段のHMMとして、予め多くの話者の発声を用いて学習した不特定話者HMMが用いられる。

【０００５】そして、認識手段104では、単語HMMを用いて 入力パターンの認識を行なう。ここで、HMMは、音声の情報源のモデルであり、音声パターンの様々な揺らぎに対処するため、統計・確率的なモデルとなっている。また、HMMの詳細な説明は、ラビナー、ジュアング著、古井訳 「音声認識の基礎(下)」、NTTアドバンステクノロジ(1995)(以下、文献1)の、102～187頁にかかれている。

【０００６】各音素のHMMは、それぞれ、通常1から10個の状態とその間の状態遷移から構成される。通常は始状態と終状態が定義されており、単位時間ごとに、各状態からシンボルが出力され、状態遷移が行なわれる。各音素の音声は、始状態から終状態までの状態遷移の間にHMMから出力されるシンボルの時系列として表される。

【０００７】各状態にはシンボルの出現確率が、また、状態間の各遷移には遷移確率が定義されている。遷移確率パラメータは音声パターンの時間的な揺らぎを表現す

るためのパラメータである。出力確率パラメータは、音声パターンの声色の揺らぎを表現するものである。始状態の確率をある値に定め、状態遷移ごとに出現確率、遷移確率を掛けていくことにより、発声がそのモデルから発生する確率を求めることができる。

【０００８】逆に、発声を観測した場合、それが、あるHMMから発生したと仮定すると、その発生確率が計算できることになる。これにより、HMMによる音声認識では、各認識候補に対してHMMを用意し、発声が入力されると、各々のHMMにおいて、発生確率を求め、最大となるHMMを 発生源と決定し、そのHMMに対応する認識候補をもって認識結果とする。

【０００９】出力確率パラメータには、離散確率分布表現と連続確率分布表現があるが、ここでは連続確率表現を例にとる。連続確率分布表現では、しばしば、混合ガウス分布、すなわち、複数のガウス分布を重みつきで加算した分布が使われる。以下の例においては、出力確率は混合ガウス連続確率分布とする。そして、出力確率パラメータ、遷移確率パラメータ、複数のガウス分布の重み係数などのパラメータは、モデルに対応する学習音声を与えて、バウムーウェルチアルゴリズムと呼ばれるアルゴリズムにより、予め学習されている。

【００１０】例えば、今、1000単語を認識対象とする場合を想定する。すなわち、1000単語の認識候補から1単 ＊語の正解を求める場合を想定する。まず、単語を認識する場合には、各音素のHMMを 連結して、認識候補単語のHMMを作成する。1000単語認識の場合には1000単語分の単語HMMを作成する。 特徴ベクトルの時系列として表現された入力パターンOを下記の（1）式により示す。

【数１】

$$O=o_1,o_2,o_3,\ldots,o_t,\ldots,o_T \quad \cdots (1)$$

ここで、Tは入力パターンの総フレーム数である。

【００１１】また、認識候補単語 W1, W2, …WNとする。ここでのNは認識候補単語数を示す。そして、各々の認識候補単語Wnの 単語HMMと、入力パターンOとの間のマッチングは、以下のように行なわれる。これからの説明においては、必要のない限り添字nを省略する。まず、単語HMMにおいて、状態jから状態iへの遷移確率をａji、出力確率分布の混合重みをｃim、各要素ガウス分布の平均ベクトルをμim、共分散行列をΣimとする。ここで、tは入力時刻、i, jはHMMの状態、mは混合要素番号を表す。前向き確率ａt(i)に関する次の漸化式計算を行う。

【００１２】この前向き確率ａt(i)は、部分的な観測系列 ｏ1, ｏ2, …, ｏtを出力し、時刻tにおいて状態iに存在する確率である。

【数２】

$$\alpha_t(i)=\pi_i \quad (i=1,2\ldots,I) \quad \cdots (2)$$

【数３】

$$\alpha_{t+1}(i)=\sum_j \alpha_t(j)\, a_{ij} b_i(o_t) \quad (i=1,\ldots,I;t=1,\ldots T) \quad \cdots (3)$$

ここで、πiは初期状態がiである確率である。

【００１３】また、(3)式におけるｂi(ｏt)は、以下に ※示す(4), (5)式により定義される。

【数４】

$$b_i(o_t)=\sum_m \lambda_{im} N(o_t;\mu_{im},\Sigma_{im}) \quad \cdots (4)$$

【数５】

$$N(o_t;\mu_{im},\Sigma_{im})=(2\pi)^{-K/2}|\Sigma_{im}|^{-1/2}\exp(-(\mu_{im}-o_t)\Sigma^{-1}(\mu_{im}-o_t)/2) \quad \cdots (5)$$

この(5)式において、Kは入力フレームおよび平均ベクトルの次元数である。

【００１４】また、単語Wnに対する入力パタンに対す ★る尤度は、以下に示す(6)式により求められる。

【数６】

$$P^n(X)=\alpha_T(I) \quad \cdots (6)$$

この(6)式において、Iは最終状態である。

【００１５】この処理を各単語モデルについて行ない、入力パターンXに対する認識結果単語Wn（ここで、下 ☆記(７)式では、nの上部に∧；ハットが付いている）は、以下に示す(7)式により求められる。

【数７】

$$\hat{n}=\mathrm{argmax}_n P^n(X) \quad \cdots (7)$$

そして、この認識結果単語Wnは、認識結果出力部に送られる。 認識結果出力部は、認識結果を画面上に出力、あるいは、認識結果に対応した制御命令を別の装置に送るなどの処理を行なう。

【００１６】次に、標準パターン作成手段102について説明する。標準パターン作成手段102は、不特定話者認識の場合、事前の多数の話者の発声を蓄積し、その発声を用いてパラメータの推定を行う。まず、以下の(8), (9)式により、後向き確率を導入する。

【数８】

$$\beta_T(i)=1 \quad (i=1,\ldots,N) \qquad \cdots (8)$$

【数９】

$$\beta_t(i)=\sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta(j), \qquad \cdots (9)$$
$$(t=T-1,T-2,\ldots,1,\ i=1,\ldots,N)$$

(9)式における$\beta t(i)$は時刻t、状態iが与えられたときの、 時刻t+1から終端までの部分的な観測系列の確率である。

【００１７】そして、前向き確率と後向き確率を用いて、観測系列Oが与えられたときに、時刻tに状態iに存在する確率は、以下に示す(10)式により与えられる。

【数１０】

$$\gamma(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{I}\alpha_t(i)\beta_t(i)} \qquad \cdots (10)$$

また、時刻tに状態iに存在し、時刻t+1に状態jに存在する 確率は、以下の(11)式により与えられる。

【数１１】

$$\zeta_{t(i,j)}=\frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{I}\sum_{j=1}^{I}\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)} \qquad \cdots (11)$$

また、混合ガウス分布の場合に、時刻t に状態番号iのk番目の 混合要素に存在する確率(占有度数)は、以下の(12)式により与えられる。

【数１２】

$$\gamma'(i,k)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{I}\alpha_t(i)\beta_t(i)}\cdot\frac{c_{ik}N(o_t,\mu_{ik},\Sigma_{ik})}{\sum_{m=1}^{M}c_{im}N(o_t,\mu_{im},\Sigma_{im})} \qquad \cdots (12)$$

【００１８】以上の計算値に基づき、 π、a、μ, Σ, cの推定値は以下の(13)～(17)式により与えられる。

【数１３】

$$\overline{\pi_i}=\gamma_1(i) \qquad \cdots (13)$$

【数１４】

$$\overline{a_{ij}}=\frac{\sum_{t=1}^{T-1}\zeta_t^{i,j}}{\sum_{t=1}^{T-1}\gamma_t^{i}} \qquad \cdots (14)$$

【数１５】

$$\overline{c_{jk}}=\frac{\sum_{t=1}^{T}\gamma_t'(j,k)}{\sum_{t=1}^{T}\gamma_t(j)} \qquad \cdots (15)$$

【数１６】

$$\overline{\mu_{jk}}=\frac{\sum_{t=1}^{T}\gamma_t'(j,k)o_t}{\sum_{t=1}^{T}\gamma_t'(j,k)} \qquad \cdots (16)$$

【数１７】

$$\overline{\Sigma_{jk}}=\frac{\sum_{t=1}^{T}\gamma_t'(j,k)(o_t-\mu_{jk})(o_t-\mu_{jk})^t}{\sum_{t=1}^{T}\gamma_t'(j,k)} \qquad \cdots (17)$$

【００１９】バウム-ウェルチアルゴリズムでは、これらの推定値をもとにパラメータを更新し、さらに、その更新されたパラメータを用いて、推定値を新たに推定するという繰り返しを行なう。そして、繰り返し毎に、観測系列の認識を行う確率が大きくなることが証明されている。以上、HMMを用いる場合を例にとり、従来の音声認識装置について説明した。

【００２０】

【発明が解決しようとする課題】さて、上述したように、出力確率分布表現には、離散分布と連続分布とがある。そして、離散分布と連続分布との中では、連続分布、特にその中でも、混合ガウス分布が、よく用いられる。この混合ガウス分布が用いられる理由は、出力確率分布表現の性能が優れているためである。

【００２１】ここで、混合ガウス分布（以下、混合分布とする）を用いる場合、その要素分布数をどの大きさにすべきかという明確な指針はない。通常は、混合分布のHMMでは、状態毎の要素分布数がすべての状態にわたり一定とし、いくつかの要素分布数を試して、その中でもっとも性能が高い要素分布数を選ぶという手続きが行われる。

【００２２】しかしながら、状態により必要な要素分布数は異なることが予想される。例えば、不必要な要素分布を多く持つとした場合、要素分布の確率を計算するための計算量の増大を招くこととなる。また、出現回数の少ない状態においては、パラメータ推定の過程で過学習が行われていまい、未知データに対する性能が劣化する可能性がある。したがって、混合分布HMMの各状態における要素分布数は、状態毎に最適化されることが望ましい。

【００２３】そして、要素分布数を状態毎に最適化する最も単純な方法は、 状態毎に要素分布数を変えて認識実験を行い、各状態毎に認識性能が高くなる要素分布数を選択する方法である。しかしながら、HMMの状態数が、通常、全体で1000から10000とたいへん多くなり、各状態毎に要素分布数を最適化することは、計算量の点

7

でほとんど不可能である。

【００２４】本発明はこのような背景の下になされたもので、混合分布を用いた確率モデルにおいて、要素分布数の調節を高速、かつ効果的に行う音声認識装置を提供することにある。

【００２５】

【課題を解決するための手段】本発明の音声認識装置は、混合分布を用いた確率モデルを用いる音声認識装置であって、標準パターンを保持する標準パターン記憶手段と、音声を入力とし標準パターンを用いて認識結果を出力する認識手段と、学習用音声を入力とし標準パターンを作成する標準パターン作成手段と、標準パターンの混合分布の要素分布数を調節する標準パターン調節手段とを具備することを特徴とする。

【００２６】本発明の音声認識装置は、混合分布を用いた確率モデルを用いる音声認識装置であって、標準パターンを保持する標準パターン記憶手段と、音声を入力とし標準パターンを用いて認識結果を出力する認識手段と、適応化用音声を入力とし標準パターンを修正する標準パターン修正手段と、標準パターンの混合分布の要素分布数を調節する標準パターン調節手段とを具備することを特徴とする。

【００２７】本発明の音声認識装置は、要素分布の木構造を作成する木構造作成手段と、学習データを入力として分布を選択する要素分布選択手段とから構成される標準パターン調節手段を具備することを特徴とする。本発明の音声認識装置は、前記標準パターン調節手段が、要素分布の選択にミニマックス法を用いるミニマックス分布選択手段を具備することを特徴とする。

【００２８】本発明の音声認識装置は、前記要素分布選択手段が、要素分布の選択において各要素分布に対応する学習データ量を選択基準として用いることを特徴とする。本発明の音声認識装置は、前記要素分布選択手段が、要素分布の選択において、記述長最小基準を選択基準として用いることを特徴とする。本発明の音声認識装置は、前記要素分布選択手段が、要素分布の選択において、赤池情報量基準を選択基準として用いることを特徴とする。

【００２９】本発明の音声認識装置は、前記木構造作成手段が、要素分布の選択において、ダイバージェンスを分布間距離として用いることを特徴とする。本発明の音声認識装置は、前記木構造作成手段が、学習データに対する尤度を分布間距離として用いることを特徴とする。本発明の音声認識装置は、混合分布を用いた確率モデルとして、隠れマルコフモデルを用いることを特徴とする。

【００３０】

ここで、αiはモデルiの次元数(自由パラメータの個数)

8

＊【発明の実施の形態】以下、図面を参照して本発明の実施形態について説明する。図１は本発明の一実施形態によるの構成を示すブロック図である。図２の従来例と異なる点は、標準パターン作成手段１０２と標準パターン記憶手段１０３との間に標準パターン作成手段２０３を挿入していることである。図１の音声認識装置のブロックにおいて、図２の音声認識装置のブロックと同様な構成（入力パターン作成手段１０２、標準パターン作成手段１０１、標準パターン記憶手段１０３、認識手段１０４）に対しては、同一の符号を付し、詳細な説明を省略する。

【００３１】この図において、入力パターン作成手段１０２は、入力される入力音声（話者の発生した音声信号）から入力パターンを作成する。また、標準パターン作成手段１０２は、従来例の説明で述べたように標準パターンを作成する。標準パターン調節手段203は、作成された標準パターンの 要素分布数を変更する。標準パターン記憶手段１０３は作成された標準パターンを記憶し、認識手段205は、入力された音声を標準パターンを用いて認識し、認識結果を出力する。

【００３２】以下に、本発明において一実施形態に、加えられた標準パターン調節手段２０３の動作について、詳細に説明する。隠れマルコフモデル（HMM）の状態における要素分布数の最適化の問題は、与えられたデータに対し最適な確率モデルを選択する問題とみなすことが可能である。この確率モデルの選択においては、過去さまざまな情報量基準が提案されてきた。

【００３３】一実施形態では、その一つであるMDL(記述長最小)を用いて分布数を最適化する方法を考える。まず、ここで上記MDLの基準について説明する。記述長最小(Minimum Description Length; MDL)基準は、最近の情報理論および計算論的学習理論の研究から、データに対し最適な確率モデルを選択する問題において、有効であることが実証されている。

【００３４】記述長最小基準は、例えば、韓太舜著、「岩波講座応用数学11、情報と符合化の数理」、 岩波書店(1994)(以下、文献2)の、249頁～275頁に説明されている。AIC(Akaike Information Criterion;赤池情報量基準)などと同様、なるべく簡単で、しかも、与えられたデータをよく 表現できるモデルが良いモデルである、という理念を具現化した基準の一つである。

【００３５】MDL基準は、確率モデルｉ＝１，…，Ｉのなかで、データｓ＝ｓ1，…，ｓNに対し、最も小さい記述長を与えるモデルを 最適なモデルとする基準である。 ここで、確率モデルiに対する記述長ＩMDL(i)は以下の(18)式で与えられる。

＊【数１８】

$$I_{MDL(i)} = -\log P_{\hat{\theta}(i)}(x^N) + \frac{\alpha_i}{2}\log N + \log I \qquad \cdots (18)$$

、 θ(i)はデータX^Nを用いて推定された モデルiの自

由パラメータ$\theta^{(i)}=(\theta_1^{(i)},\cdots,\theta\alpha_i^{(i)})$の最尤推定量である。

【００３６】上記(18)式において、第1項はデータに対する対数尤度(以下、尤度と記す)に負符号を付けた量であり、第2項はモデルの複雑さを表す量であり、第3項はモデルiを選択するために要する記述長である。このように、モデルがより複雑なほど、データに対する尤度が大きくなり、したがって 第1項の値は減少する。一方、モデルが複雑になれば、自由パラメータ数が増加するため、第2項の値は増加する。このように、第1項と第2項の間にはトレードオフの関係があり、記述長ＩMDL(i)は、適当な複雑さを有するモデルで最小値をとることが期待される。

【００３７】そして、このMDL基準を用いた状態毎の要素分布数最適化アルゴリズムは、以下の通りである。まず、通常の手順で学習データを用いた混合ガウス分布HMMの学習を行う。この際、要素分布数は全状態にわたり一定とし、上限と考えられる数まで、要素分布数を増やしたHMMを学習する。また、学習の過程で要素分布ごとの占有度数$\gamma't(i,k)$を保存しておく。ここでiは状態の添字、kは状態における要素分布の添字である。

【００３８】次に、標準パターン調整手段２０３は、各状態において要素分布数の最適化を行う。なお、この先は一つの状態iについてのみ説明することにし、状態の添字iを省略する。標準パターン調整手段２０３は、他の状態に対しても同じ処理を行う。まず、標準パターン調整手段２０３は、内部の木構造作成手段により、状態＊

＊ごとに要素分布の木構造を作成する。ここで、ルートは一つの分布であり、リーフは各要素分布である。

【００３９】このとき、要素分布の木構造を作成するには様々な方法が考えられるが、ここでは2分木をk-meansアルゴリズムを用いて作成する。また、各要素分布間の距離（分布間距離）としては、カルバックダイバージェンスを用いる。このカルバックダイバージェンスは、ガウス分布の平均・共分散の値から容易に計算できる。この要素分布の木構造作成方法については、特許第００２５３１０７３号，上記文献２に詳細に記載されている。

【００４０】次に、標準パターン調整手段２０３は、上記木構造の各ノードの分布(ノード分布)の分散を求める。ここで、各ノード分布の分散は、その支配するすべてのリーフの要素分布の占有度数とガウス分布パラメータから求められる。今、この木構造を上下に分断するノード分布の集合を「カット」と呼ぶ。このカットの数は多数あるが、一つ一つのカットが、その状態における一つの確率モデルとなる。ここで、MDL基準を用いて最適なカットを求めることを考える。

【００４１】例えば、あるカットUに対する記述長は次のように計算される。ここで、カットUを構成するノード分布を Ｓ1, …ＳMとする。 ここで、Mはカットにおけるノード分布の個数である。これにより、データの分布Ｓmに対する尤度Ｌ(Ｓm)は以下に示す（19），（20）式のように近似できる。

【数１９】

$$L(s_m)=\sum_{t=1}^{T}\sum_{s\in S_m}\log(N(o_t,\mu_{Sm},\Sigma_{Sm}))\gamma_{t(s)}$$

$$=-\frac{1}{2}(\log((2\pi)^K|\Sigma|+K)\Gamma_{(Sm)} \quad \cdots(19)$$

【００４２】上記（１９）式において、

【数２０】

$$\Gamma_{(Sm)}=\sum_{t=1}^{T}\sum_{s\in S_m}\gamma_{t(s)} \quad \cdots(20)$$

であり、sは分布Ｓmの下のすべてのリーフ分布であり、Kは共有標準パターンとして用いられる平均ベクトルおよび分散の次元数である。また、（19）式において、μ※

※Sm, ΣSmは、それぞれ分布Smにおける平均ベクトルおよび分散である。

【００４３】上述した結果を用いることにより、カットUに対する記述長Ｉ(U)は、以下の(21)式のように記述することができる。

【数２１】

$$I(U)=\sum_{m=1}^{M}L(Sm)+\frac{1}{2}\cdot 2KM\log\sum_{m=1}^{M}\Gamma(Sm)$$

$$=\frac{1}{2}\sum_{m=1}^{M}\Gamma(Sm)\log(|\Sigma Sm|)+KM\log V+\frac{K}{2}(1+\log(2\pi))\cdot V$$

$$\cdots(21)$$

ここで、

【数２２】

$$V=\sum_{m=1}^{M}\Gamma(Sm) \quad \cdots(22)$$

であり、このVはUに対応するすべてのデータのフレーム数に相当する量であり、分割の方法によらず一定値である。

【００４４】そして、標準パターン調整手段２０３は、すべての可能なカットに関して、記述長 I(U)を計算し、最も小さいＩ(U)をもつカットUを選択する。このとき、可能な分割の種類、すなわち、カットUの数は通常大変多くなる。そこで、次のようなアルゴリズムを用い

ることにより、カットUの選択時の計算量を節約する。以下、ある状態pの要素分布数最適化について述べる。

【００４５】まず、状態pに対するノード(節点)を作成する。ここで、このノードをルートノードと呼ぶ。ルートノードの分布パラメータは、この状態に対応するすべての要素分布に対応するすべてのデータサンプルから推定される。例えば、木構造が2分木であり、ルートノードの分布をS0、その2つの子ノードの分布をS1、S2としたとき、親ノードから子ノードへ展開したときの記述長の変化分は以下の(23)式で記述される。

【数２３】

$$\Delta = l(S_1,S_2) - l(S_0) = \frac{1}{2}\left(\Gamma(S_1)\log|\Sigma_{S_1}| + \Gamma(S_2)\log|\Sigma_{S_2}| - \Gamma(S_0)\log|\Sigma_{S_0}|\right) + K\log V \quad \cdots (23)$$

【００４６】例えば、標準パターン調整手段２０３は、Δ＜０である場合、親ノードの展開を行い、一方、Δ＞０である場合、親ノードの展開を行わない。また、展開するときには、さらに子ノードS1、S2それぞれについて、上述した処理と同様に、その子ノードへ展開したときの記述長の変化を計算し、展開するか否かを判断するという処理を繰り返す。そして、すべてのノードの展開が終ったとき、その展開の末端のノードの集合がカットとなり、そのノード分布が要素分布として選択されたことになる。そして、改めて選択された分布のみを要素分布としてもつ、混合ガウス分布HMMを作成し、その要素分布を改めて学習におけるデータにより学習する手続きを行う。

【００４７】以上が、図１に示す一実施形態の音声認識装置の説明である。ここでは、隠れマルコフモデル（HMM）を例にして説明したが、モデルが混合ガウス分布である場合にも容易に適用可能である。これは、請求項１０の発明に対応している。また、上述した一実施形態の説明では、音響モデル学習について説明したが、使用者の少量の発声を用いて標準パターンの修正を行うような、話者適応を行う際にも、話者適応用データを用いて要素分布数の調節を行うことが可能である。この場合、発明の音声認識装置の構成としては、標準パターン作成手段のかわりに、標準パターン修正手段を用い、この標準パターン修正手段への入力音声は、認識用の入力パターン作成手段に用いる話者と同一の話者の音声を用いる。

【００４８】また、上述した一実施形態の音声認識装置においては、木構造による要素分布数の調節手段について説明したが、ミニマックス法を用いたミニマックス分布選択手段による調節も、以下のように行うことができる。以下、一つの状態について説明する。まず、学習データ中にある回数(X回)以上、出現した分布の集合をAとし、そうでない分布をBとする。Aに属する分布とBに属する分布とのの距離をすべて計算し、Bの分布のうち、最も近いAの分布からの距離が最も大きい分布を取り除く。

【００４９】次に、その分布以外のBの分布のうち、最も近いAの分布からの距離が最も大きい分布を取り除く。この手続きを分布数が予め定めた最小分布数になるまで繰り返す。そして、最小分布数より小さくならない（すなわち、Bの分布数が小さい）ときには、その時点で上述の処理を停止する。以上は、請求項４の発明に対応する。

【００５０】また、一実施形態においては、ノードの選択にMDL基準を用いたが、データ量閾値を用いることも可能である。すなわち、データ量が有る閾値以上ある分布のうちもっともリーフに近い分布の集合をカットとする。以上は、請求項５の発明に対応する。

【００５１】さらに、一実施形態においては、情報量基準としてMDL基準を用いる場合についてのみ説明したが、赤池情報量基準(AIC)を用いた場合、あるいは他の類似の情報量基準を用いた場合においても容易に適用可能である。以上は、請求項７の発明に対応する。

【００５２】加えて、一実施形態においては、ダイバージェンスを分布間の距離として用いたが、分布を共有したときの尤度の増分を距離値として用いることもできる。以上は、請求項９の発明に対応する。

【００５３】以上、本発明の一実施形態を図面を参照して詳述してきたが、具体的な構成はこの実施形態に限られるものではなく、本発明の要旨を逸脱しない範囲の設計変更等があっても本発明に含まれる。

【００５４】

【発明の効果】本発明の音声認識装置によれば、新たに加えたパラメータ調節手段を用いて、混合ガウス分布を用いたパターン認識において、音声の標準パターンの要素分布数を、HMMの状態毎に要素分布数を最適化、すなわち、HMMの状態毎に認識性能が高くなる要素分布数に調節することにより、不必要な要素分布を省くことができ、過学習による未知の音声データに対する劣化を防止することとなり、高性能な音声認識を行うことが可能になる。

【図面の簡単な説明】

【図１】　本発明の一実施形態による音声認識装置の構成を示すブロック図である。

【図２】　従来例による音声認識装置の構成を示すブロック図である。

【符号の説明】

１０１　入力パターン作成手段

１０２　標準パターン作成手段

13

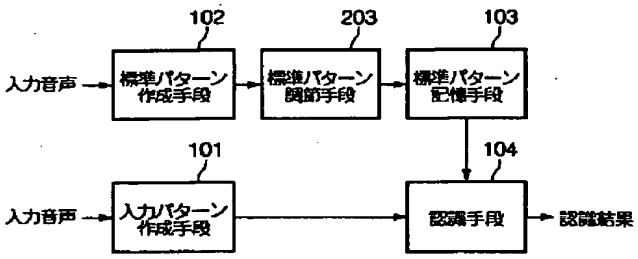
１０３　標準パターン記憶手段

１０４　認識手段

14

２０３　標準パターン調節手段

【図１】



【図２】